Data Science with R

By Anthony Castillo

What is R?

https://www.r-project.org

https://www.rstudio.com

https://www.minsalud.gov.co/salud/POS/Paginas/unidad-de-pago-por-capitacion-upc.aspx

Our data…

Click on the “2018” link (Under “Estudios de Suficiencia”) and scroll down to the bottom of page 79 for this lecture’s data.

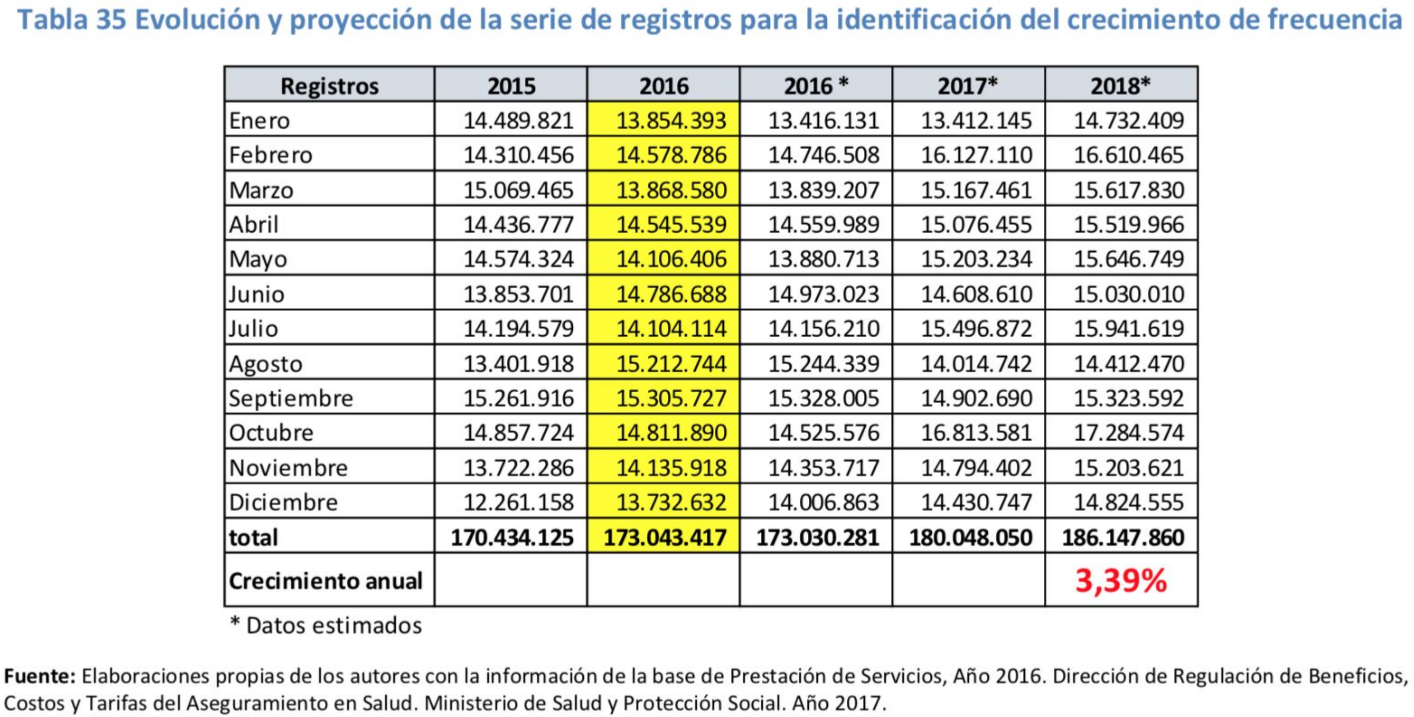


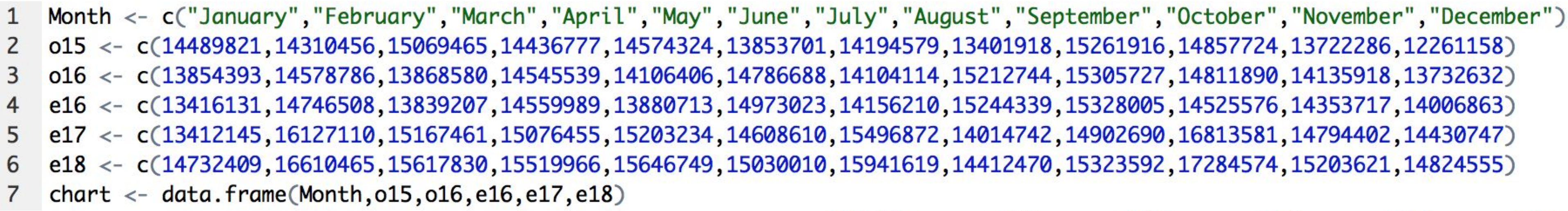
Tabla 35 Evolución y proyección de la serie de registros para la identificación del crecimiento de frecuencia

Registros	2015	2016	2016 *	2017*	2018*
Enero	14.489.821	13.854.393	13.416.131	13.412.145	14.732.409
Febrero	14.310.456	14.578.786	14.746.508	16.127.110	16.610.465
Marzo	15.069.465	13.868.580	13.839.207	15.167.461	15.617.830
Abril	14.436.777	14.545.539	14.559.989	15.076.455	15.519.966
Mayo	14.574.324	14.106.406	13.880.713	15.203.234	15.646.749
Junio	13.853.701	14.786.688	14.973.023	14.608.610	15.030.010
Julio	14.194.579	14.104.114	14.156.210	15.496.872	15.941.619
Agosto	13.401.918	15.212.744	15.244.339	14.014.742	14.412.470
Septiembre	15.261.916	15.305.727	15.328.005	14.902.690	15.323.592
Octubre	14.857.724	14.811.890	14.525.576	16.813.581	17.284.574
Noviembre	13.722.286	14.135.918	14.353.717	14.794.402	15.203.621
Diciembre	12.261.158	13.732.632	14.006.863	14.430.747	14.824.555
total	**170.434.125**	**173.043.417**	**173.030.281**	**180.048.050**	**186.147.860**
Crecimiento anual					**3,39%**

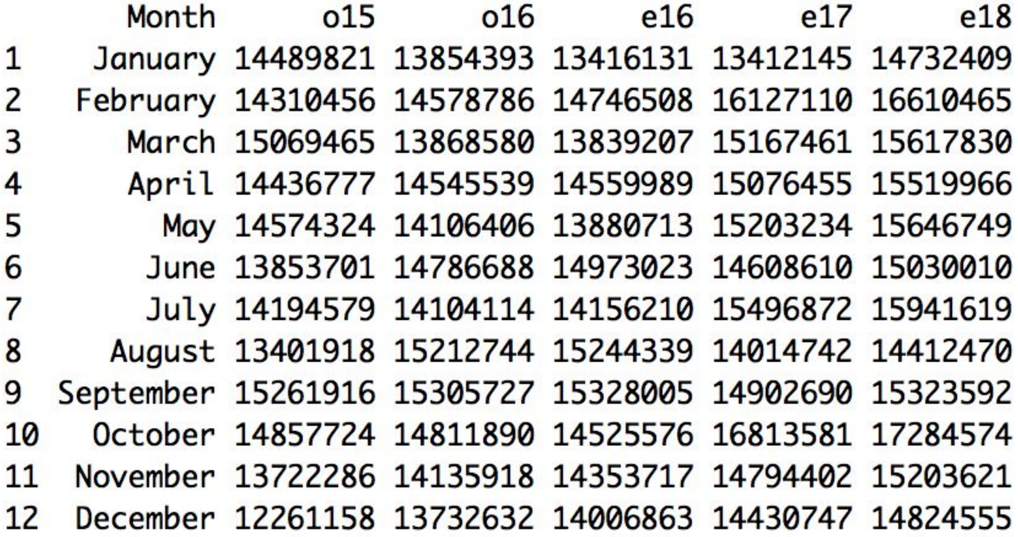
* Datos estimados

Fuente: Elaboraciones propias de los autores con la información de la base de Prestación de Servicios, Año 2016. Dirección de Regulación de Beneficios, Costos y Tarifas del Aseguramiento en Salud. Ministerio de Salud y Protección Social. Año 2017.

Creating the Dataframe…

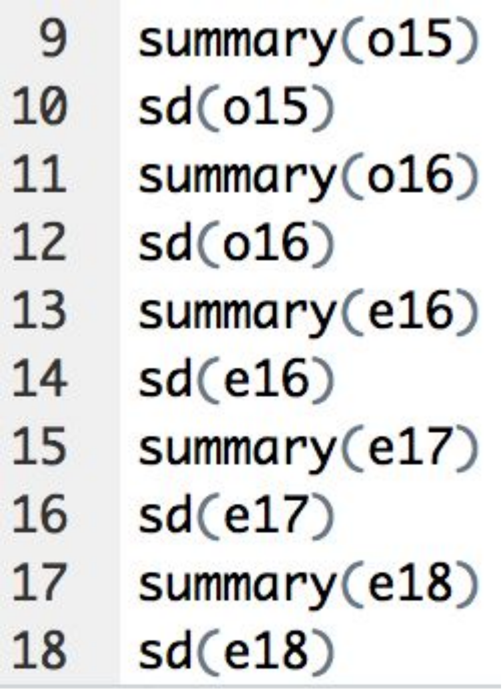


```
Month <- c("January","February","March","April","May","June","July","August","September","October","November","December")
o15 <- c(14489821,14310456,15069465,14436777,14574324,13853701,14194579,13401918,15261916,14857724,13722286,12261158)
o16 <- c(13854393,14578786,13868580,14545539,14106406,14786688,14104114,15212744,15305727,14811890,14135918,13732632)
e16 <- c(13416131,14746508,13839207,14559989,13880713,14973023,14156210,15244339,15328005,14525576,14353717,14006863)
e17 <- c(13412145,16127110,15167461,15076455,15203234,14608610,15496872,14014742,14902690,16813581,14794402,14430747)
e18 <- c(14732409,16610465,15617830,15519966,15646749,15030010,15941619,14412470,15323592,17284574,15203621,14824555)
chart <- data.frame(Month,o15,o16,e16,e17,e18)
```

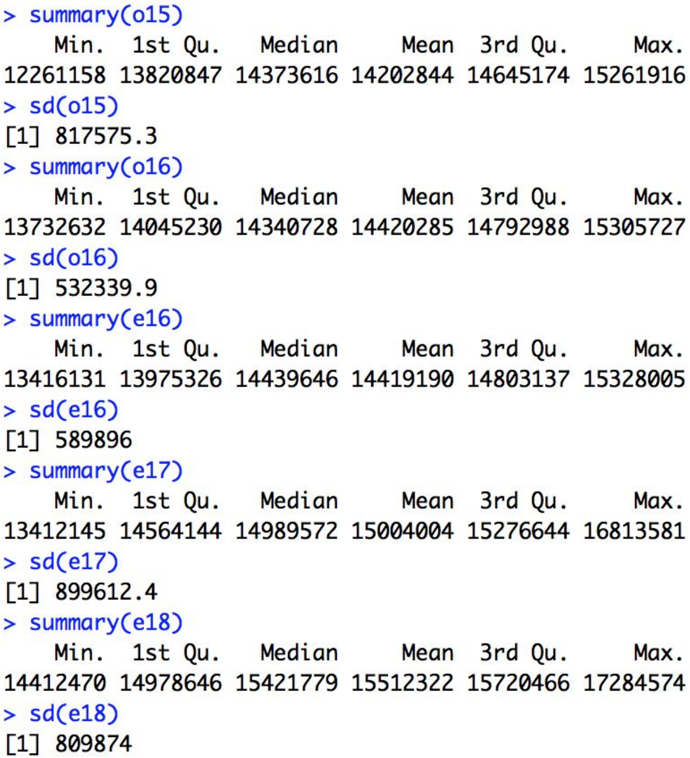


```
       Month      o15      o16      e16      e17      e18
1    January 14489821 13854393 13416131 13412145 14732409
2   February 14310456 14578786 14746508 16127110 16610465
3      March 15069465 13868580 13839207 15167461 15617830
4      April 14436777 14545539 14559989 15076455 15519966
5        May 14574324 14106406 13880713 15203234 15646749
6       June 13853701 14786688 14973023 14608610 15030010
7       July 14194579 14104114 14156210 15496872 15941619
8     August 13401918 15212744 15244339 14014742 14412470
9  September 15261916 15305727 15328005 14902690 15323592
10   October 14857724 14811890 14525576 16813581 17284574
11  November 13722286 14135918 14353717 14794402 15203621
12  December 12261158 13732632 14006863 14430747 14824555
```

Getting the Number Summaries...



```
summary(o15)
sd(o15)
summary(o16)
sd(o16)
summary(e16)
sd(e16)
summary(e17)
sd(e17)
summary(e18)
sd(e18)
```

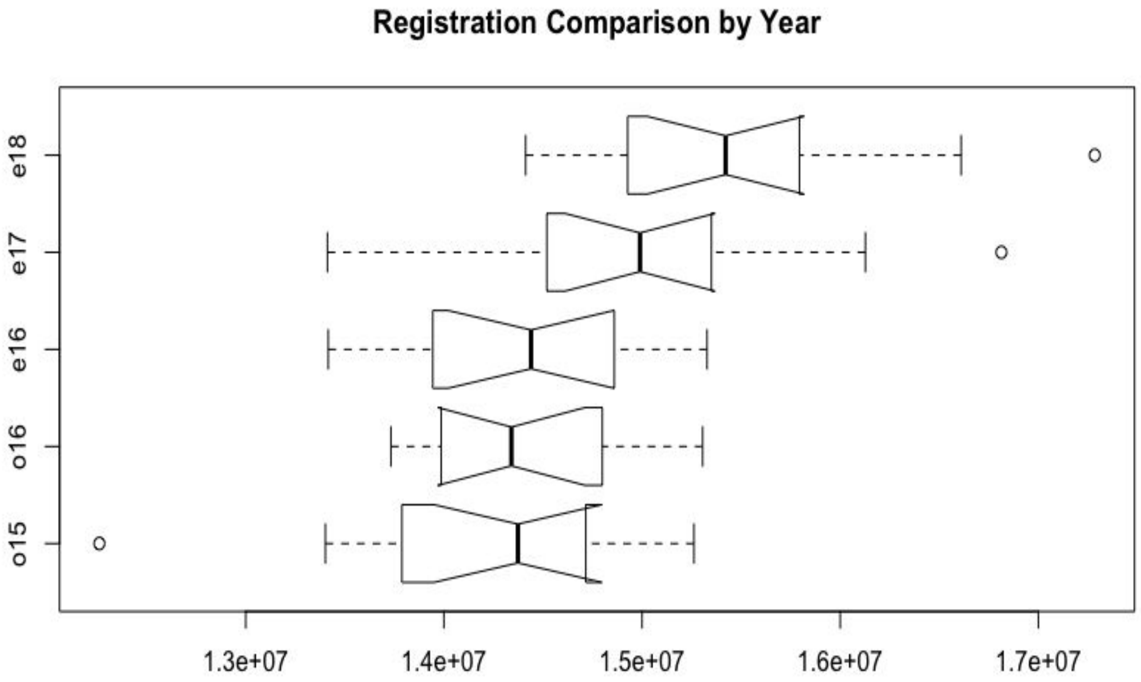


```
> summary(o15)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
12261158 13820847 14373616 14202844 14645174 15261916 
> sd(o15)
[1] 817575.3
> summary(o16)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
13732632 14045230 14340728 14420285 14792988 15305727 
> sd(o16)
[1] 532339.9
> summary(e16)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
13416131 13975326 14439646 14419190 14803137 15328005 
> sd(e16)
[1] 589896
> summary(e17)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
13412145 14564144 14989572 15004004 15276644 16813581 
> sd(e17)
[1] 899612.4
> summary(e18)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
14412470 14978646 15421779 15512322 15720466 17284574 
> sd(e18)
[1] 809874
```

Creating the Box-and-Whisker Plots:

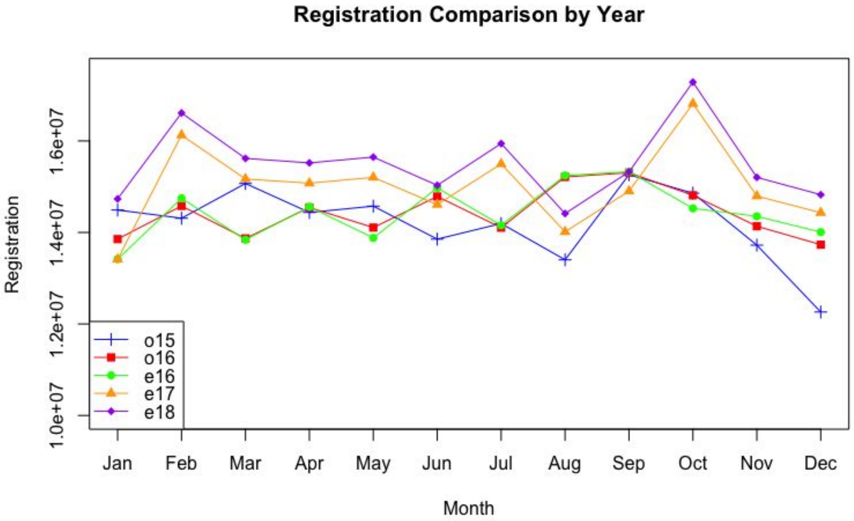
19)

```
boxplot(o15,o16,e16,e17,e18,
main="Registration Comparison by Year",
names=c("o15","o16","e16","e17","e18"),
horizontal=TRUE, notch=TRUE)
```

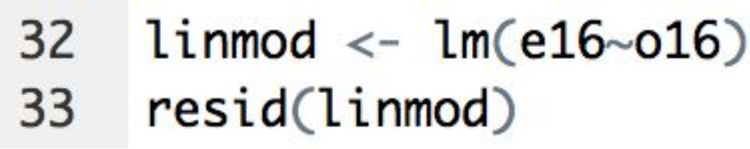


Creating the Line Plots...

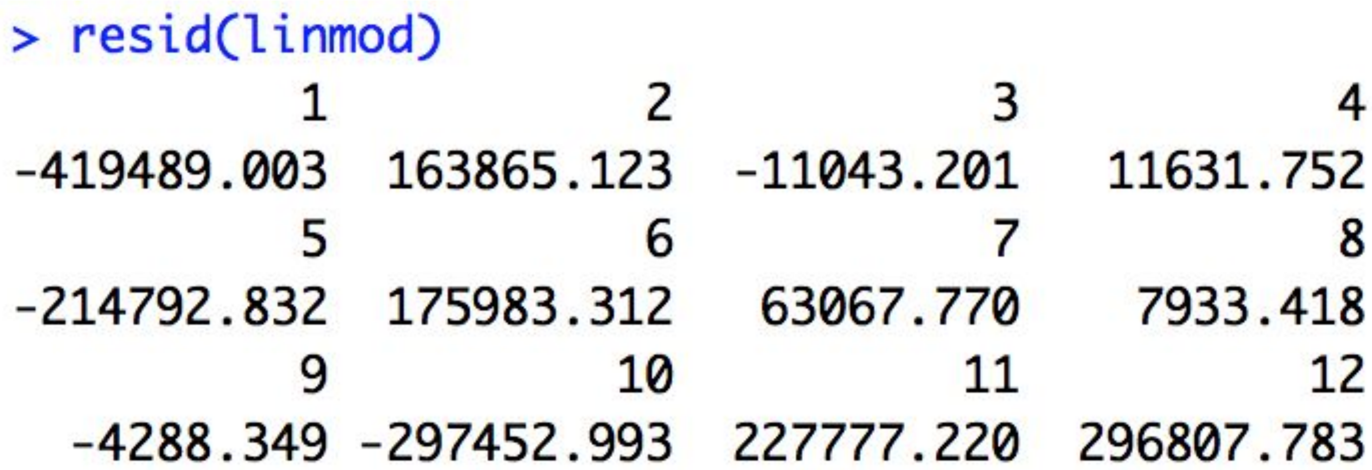
```r
plot(x=c(1:12),o15,type="o",col="blue",pch=3,lty=1,ylim=c(10000000,17500000),main="Registration Comparison by Year",xlab="Month",ylab="Registration",xaxt='n')
axis(side=1, at=c(1:12), labels=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))
points(x=c(1:12),o16,col="red",pch=15)
lines(x=c(1:12),o16,col="red",lty=1)
points(x=c(1:12),e16,col="green",pch=16)
lines(x=c(1:12),e16,col="green",lty=1)
points(x=c(1:12),e17,col="orange",pch=17)
lines(x=c(1:12),e17,col="orange",lty=1)
points(x=c(1:12),e18,col="purple",pch=18)
lines(x=c(1:12),e18,col="purple",lty=1)
legend("bottomleft",legend=c("o15","o16","e16","e17","e18"),
       col=c("blue","red","green","orange","purple"), pch=c(3,15,16,17,18), lty=1)
```



Calculating the 2016 Residuals...

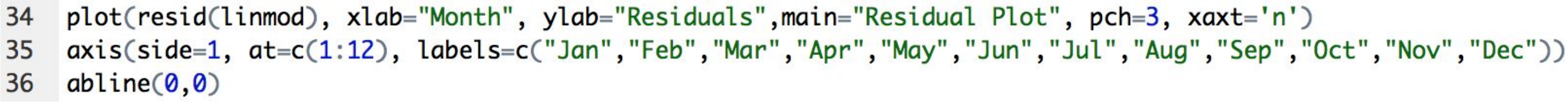


```r
linmod <- lm(e16~o16)
resid(linmod)
```

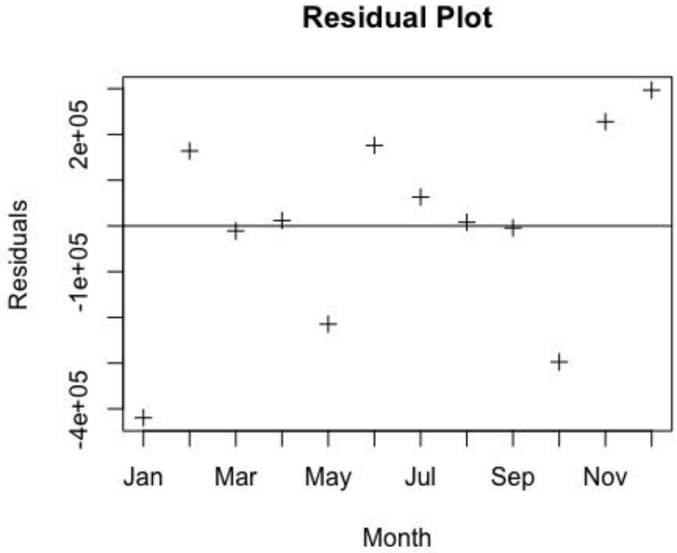


```
> resid(linmod)
          1           2           3           4 
-419489.003  163865.123  -11043.201   11631.752 
          5           6           7           8 
-214792.832  175983.312   63067.770    7933.418 
          9          10          11          12 
  -4288.349 -297452.993  227777.220  296807.783 
```

Plotting the Residuals

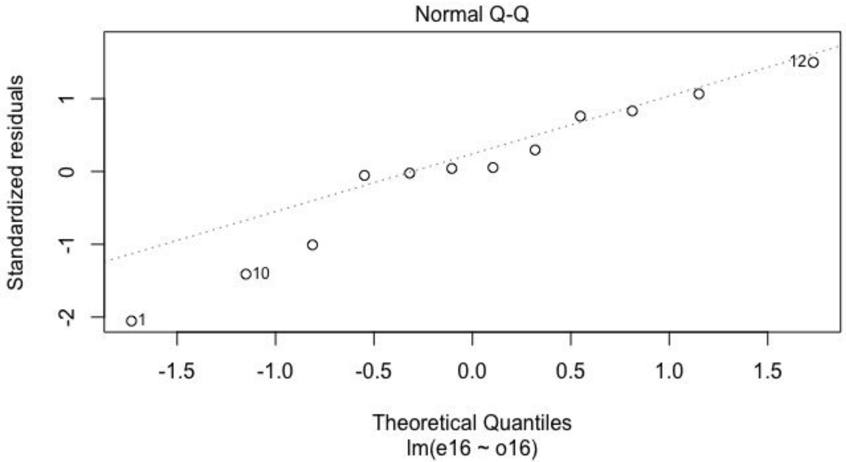


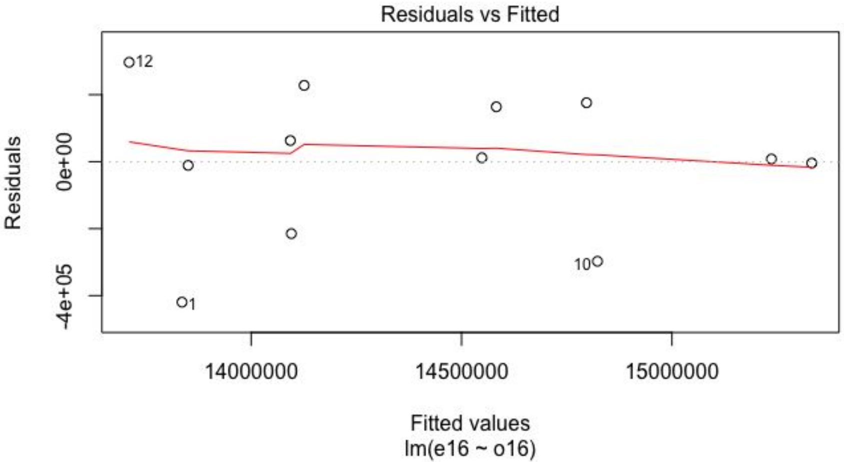
```
plot(resid(linmod), xlab="Month", ylab="Residuals",main="Residual Plot", pch=3, xaxt='n')
axis(side=1, at=c(1:12), labels=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))
abline(0,0)
```



Analyzing the Residual Plots…

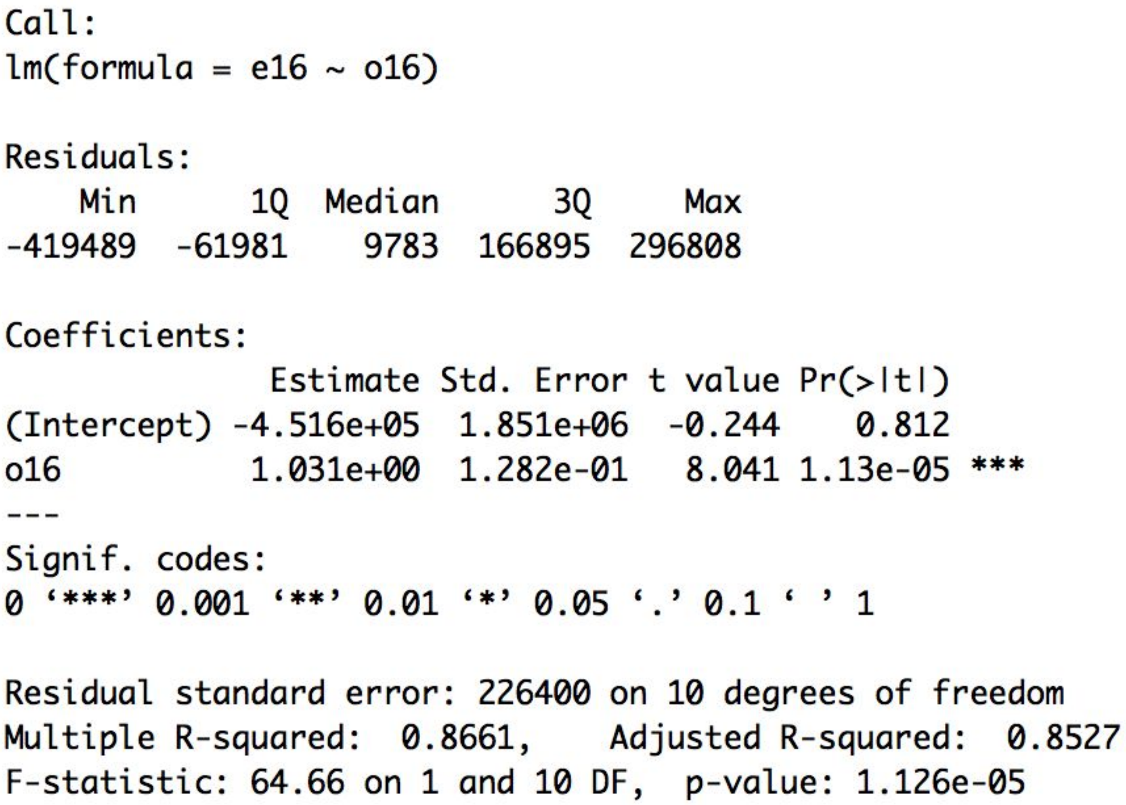
37) plot(linmod)





Interpreting the Residual Number Summary…

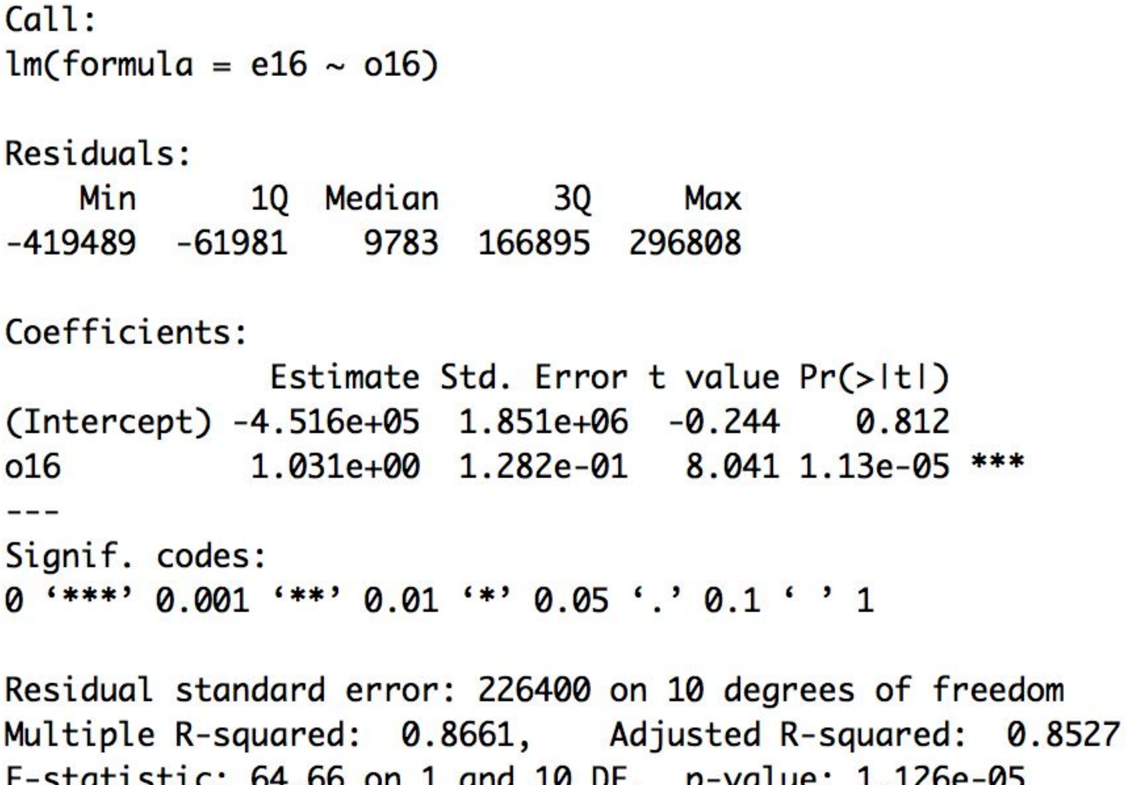
38) summary(linmod)



```
Call:
lm(formula = e16 ~ o16)

Residuals:
    Min      1Q  Median      3Q     Max 
-419489  -61981    9783  166895  296808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.516e+05  1.851e+06  -0.244    0.812    
o16          1.031e+00  1.282e-01   8.041 1.13e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226400 on 10 degrees of freedom
Multiple R-squared:  0.8661,	Adjusted R-squared:  0.8527 
F-statistic: 64.66 on 1 and 10 DF,  p-value: 1.126e-05
```

Deriving the Normal Linear Regression Coefficients…

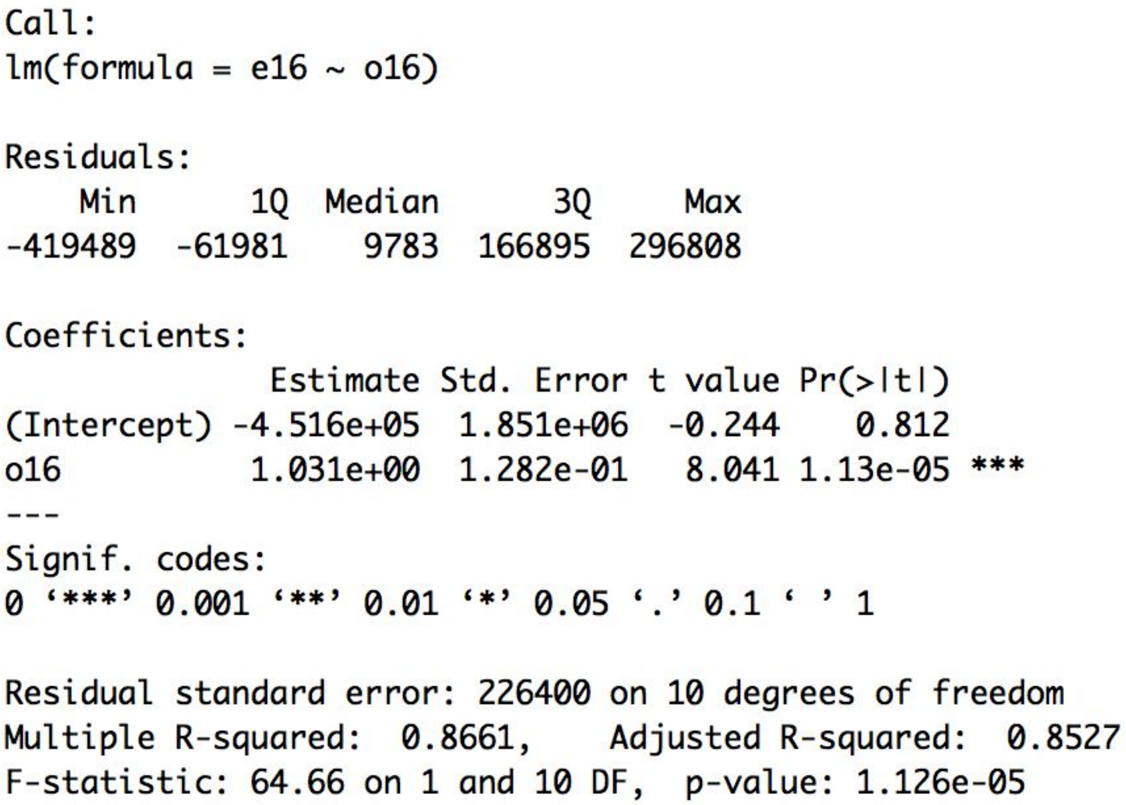
38) summary(linmod)



```
Call:
lm(formula = e16 ~ o16)

Residuals:
    Min      1Q  Median      3Q     Max 
-419489  -61981    9783  166895  296808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.516e+05  1.851e+06  -0.244    0.812    
o16          1.031e+00  1.282e-01   8.041 1.13e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226400 on 10 degrees of freedom
Multiple R-squared:  0.8661,	Adjusted R-squared:  0.8527 
F-statistic: 64.66 on 1 and 10 DF,  p-value: 1.126e-05
```

Analyzing the R-squared Coefficient…

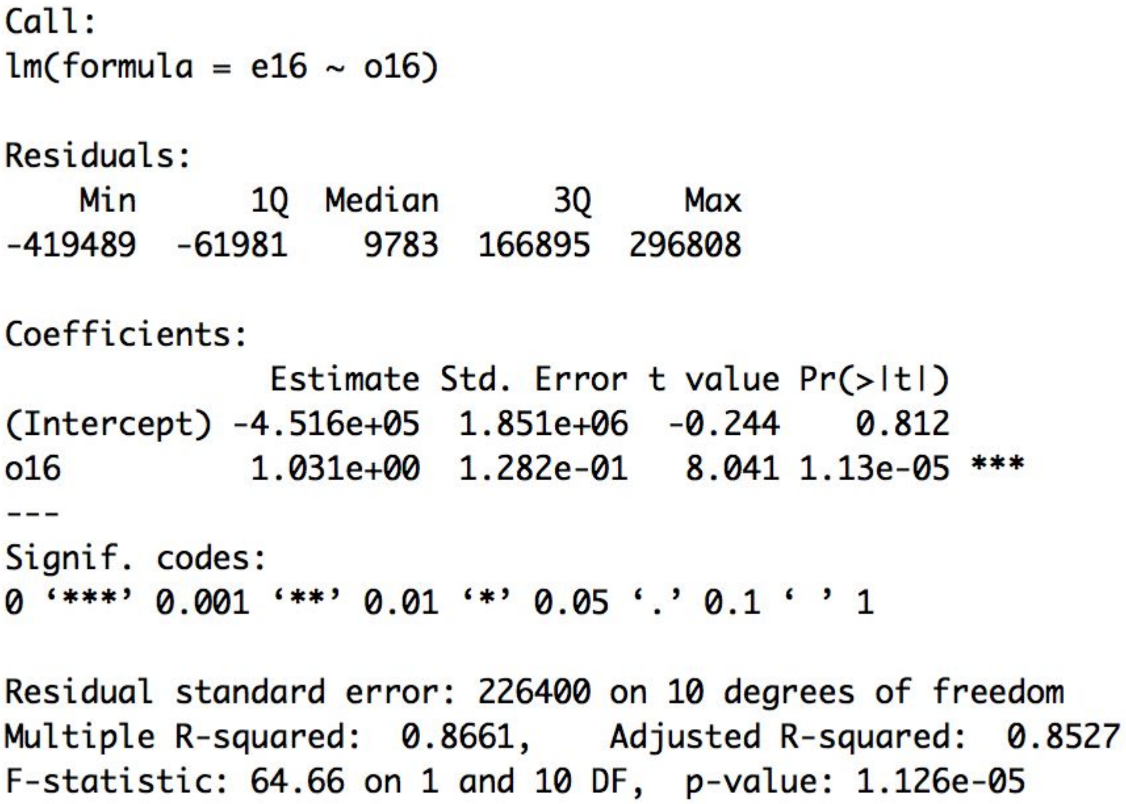
38) summary(linmod)



```
Call:
lm(formula = e16 ~ o16)

Residuals:
    Min      1Q  Median      3Q     Max 
-419489  -61981    9783  166895  296808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.516e+05  1.851e+06  -0.244    0.812    
o16          1.031e+00  1.282e-01   8.041 1.13e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226400 on 10 degrees of freedom
Multiple R-squared:  0.8661,	Adjusted R-squared:  0.8527 
F-statistic: 64.66 on 1 and 10 DF,  p-value: 1.126e-05
```

Testing the Null Hypothesis…

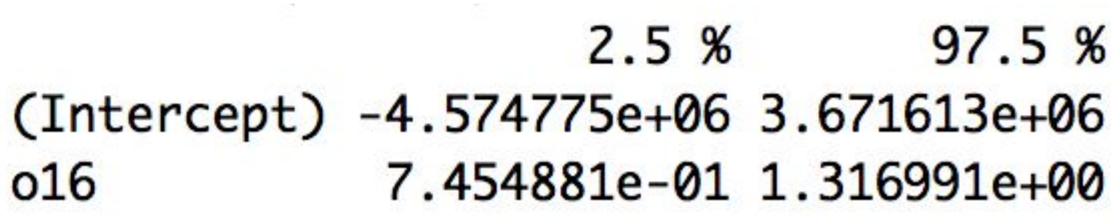
38) summary(linmod)



```
Call:
lm(formula = e16 ~ o16)

Residuals:
    Min      1Q  Median      3Q     Max 
-419489  -61981    9783  166895  296808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.516e+05  1.851e+06  -0.244    0.812    
o16          1.031e+00  1.282e-01   8.041 1.13e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226400 on 10 degrees of freedom
Multiple R-squared:  0.8661,	Adjusted R-squared:  0.8527 
F-statistic: 64.66 on 1 and 10 DF,  p-value: 1.126e-05
```

Deriving the 95% Confidence Interval...

39) confint(linmod)



```
                    2.5 %       97.5 %
(Intercept) -4.574775e+06 3.671613e+06
o16          7.454881e-01 1.316991e+00
```

Why?

Sources...

PSTAT 10 w/ Prof. Dawn Holmes (Fall 2017)

PSTAT 126 w/ Prof. Todd Gross (Spring 2018)

Any questions?